# Initial Performance Evaluation of the NetEffect 10 Gigabit iWARP Adapter

**Dennis Dalessandro**
**Ohio Supercomputer Center**

# Who is behind this?



Dennis Dalessandro
dennis@osc.edu

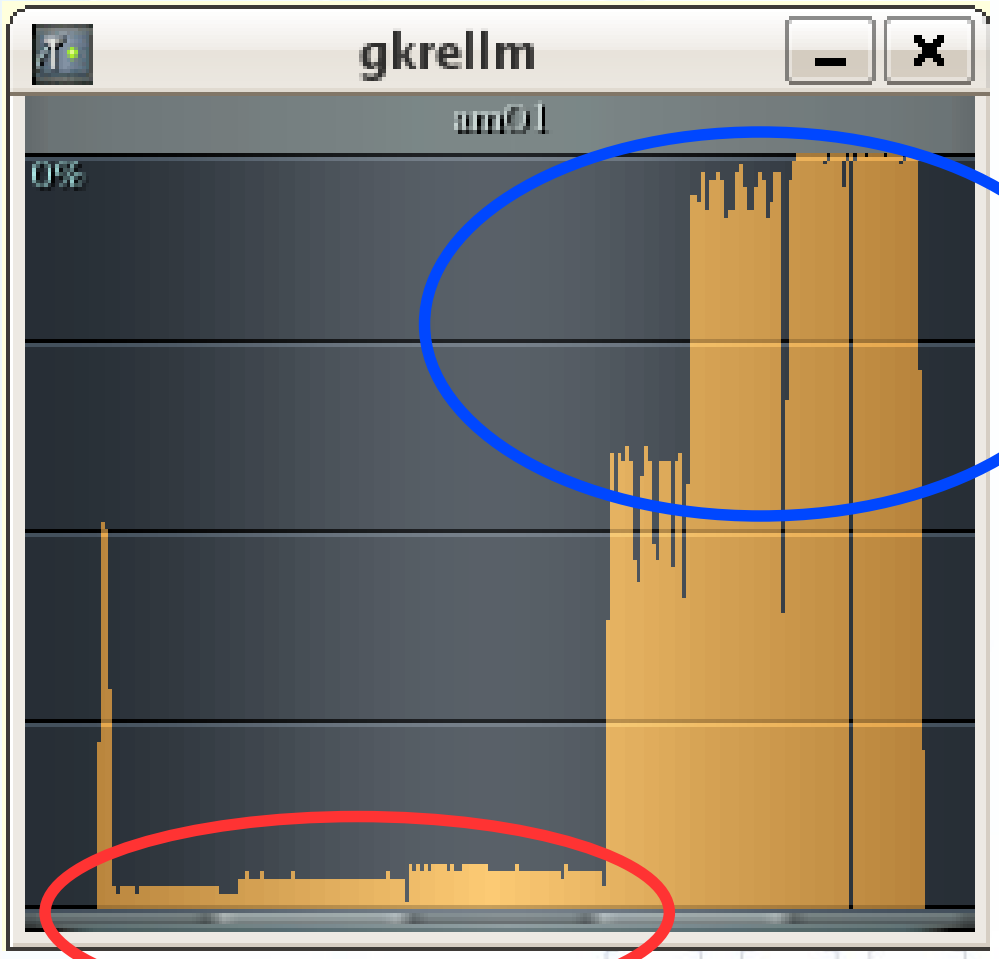

Pete Wyckoff
pw@osc.edu


Gary Montry
gmontry@neteffect.com


# Background

- RDMA
  - Zero-Copy (OS Bypass)
  - Protocol Offload
- Examples of RDMA
  - InfiniBand
  - Myrinet
  - iWARP (What this talk is about)
- Why?
  - Processing network stack is expensive (TCP)

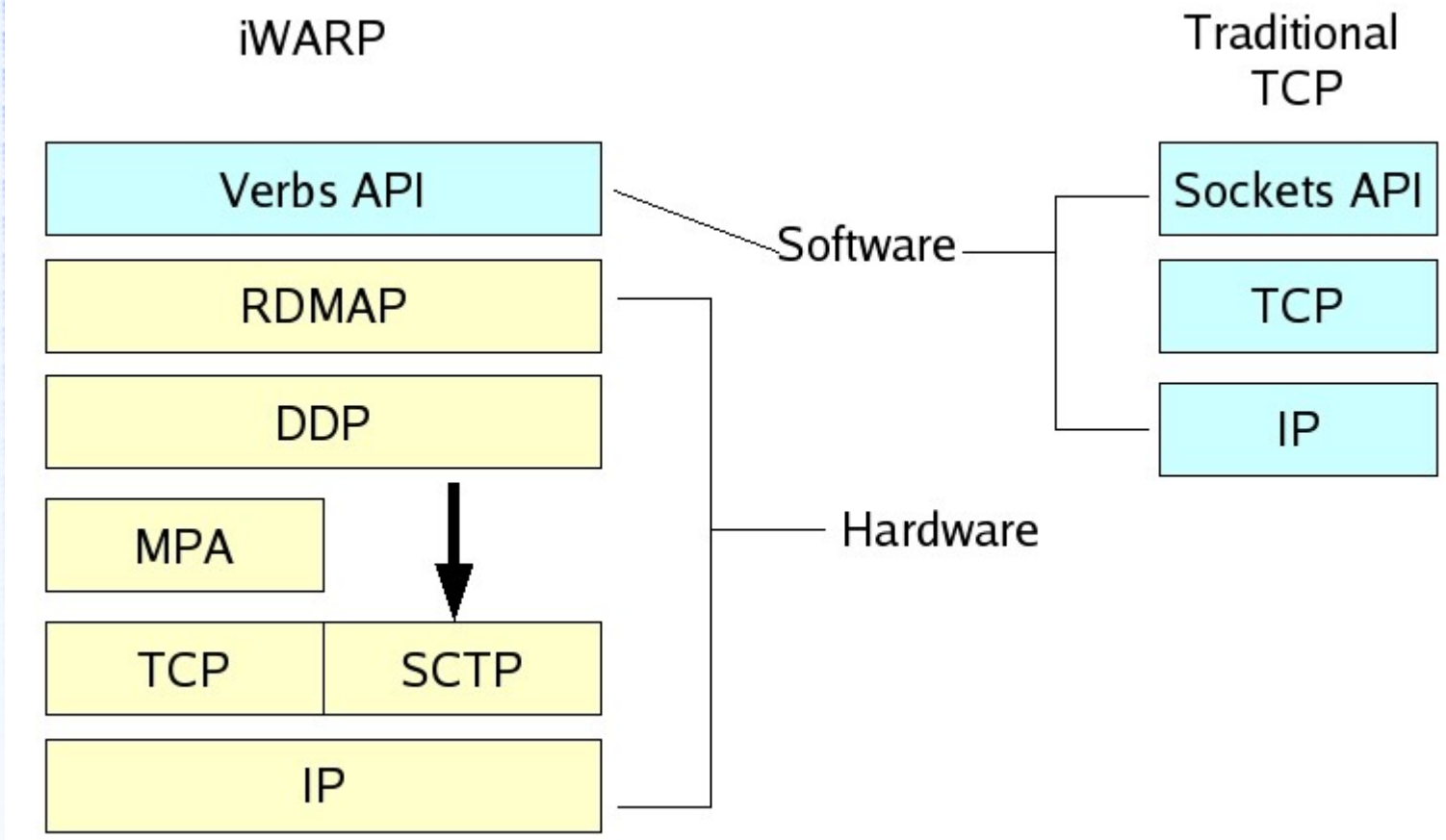
# CPU Usage



# What is iWARP?

- RDMA over TCP
  - aka RDMA over Ethernet
- IETF Specifications
  - RDMAP
  - DDP
  - MPA
- All the good things of IB over a commodity transport

# iWARP layers



# iWARP as an alternative to IB

- iWARP works over TCP/IP
  - The world runs on TCP (Ethernet)
    - Nothing IB vendors can do to change this
  - iWARP works natively in the WAN
- Downside
  - Expense of 10 Gig hardware
    - History of Ethernet tells us price will drop rapidly
    - Switch cost has already decreased greatly
    - Adapter cost will fall as demand increases
  - IB higher throughput with latest and greatest DDR
  - IB lower latency

# What iWARP is not.....

- NOT an alternative to Ethernet
  - iWARP is the next feature of Ethernet/TCP
  - Eventually integrated on-board
- NOT a new idea
  - RDMA has been long utilized
  - Ethernet is ubiquitous
  - Simply combines them
- NOT just for HPC
  - Attractive for many uses and environments
    - High demand web servers
    - Storage

OSC

# What about TOE?

- TOE is TCP Offload Engine
  - Offloads protocol processing to network device
  - Huge performance help
  - Not a full solution
    - Lacks Zero-Copy thus higher CPU usage
    - Moving memory and making copies is a HUGE cost

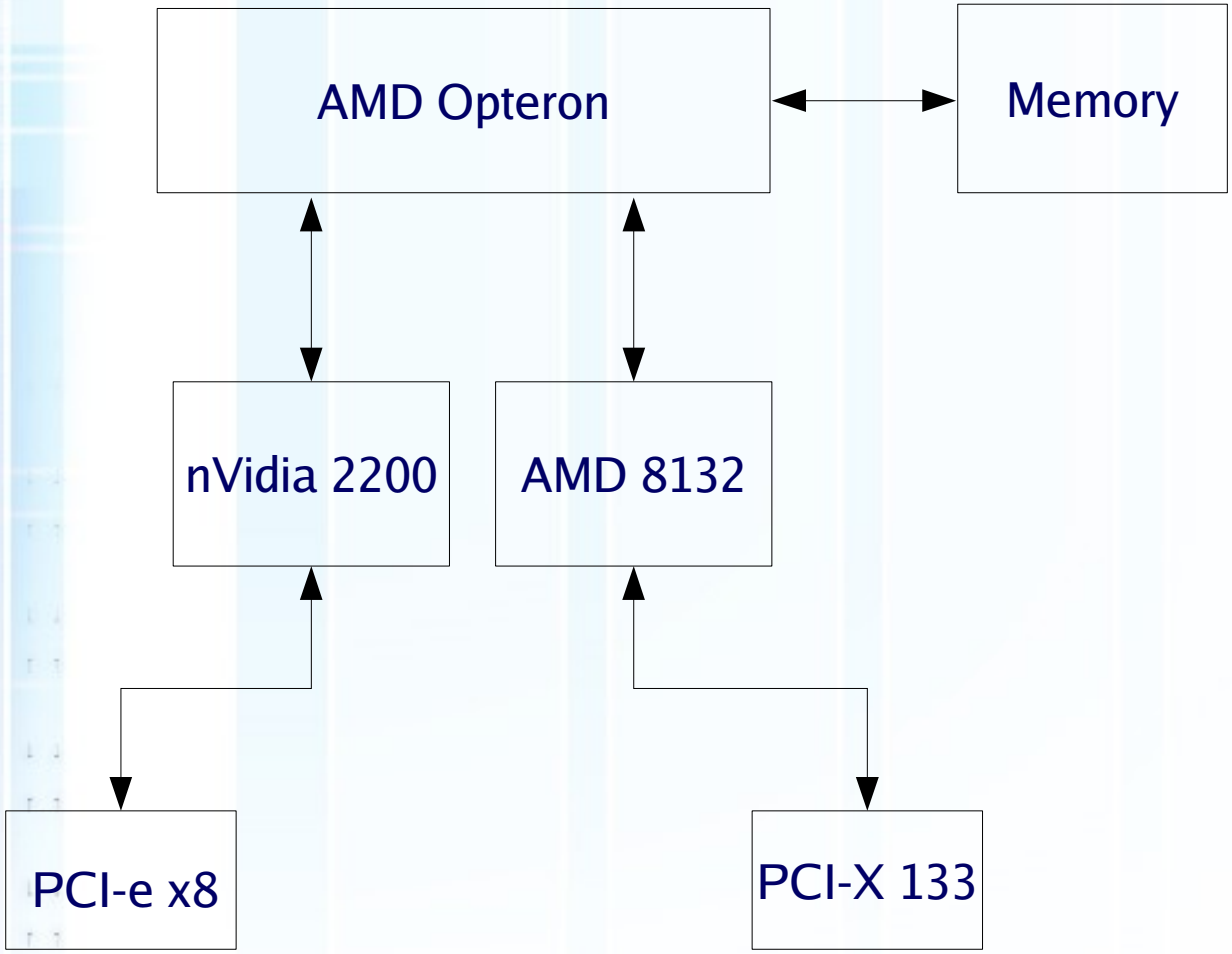
# iWARP Hardware

- Started with 1 Gig adapter from Ammasso
  - Closed up shop 9/05
- NetEffect first available 10 Gig Adapter!
  - Supports OpenFabrics, though not in SVN yet
- Chelsio
  - OpenFabrics driver but HW not generally available
- Other companies?

# NetEffect 10 Gigabit iWARP Adapter

- PCI-X Interface
  - Throughput limited by PCI bus
  - Full 10Gig possible with upcoming PCIe
- Beats 4X IB in terms of throughput
  - experimental results in the following slides
- Unfair to compare to IB
  - Vendors complain about not using DDR IB
  - DDR is 20Gbps and beyond
  - 4X IB marketed as 10Gig but really 8Gig
  - iWARP really is 10Gig

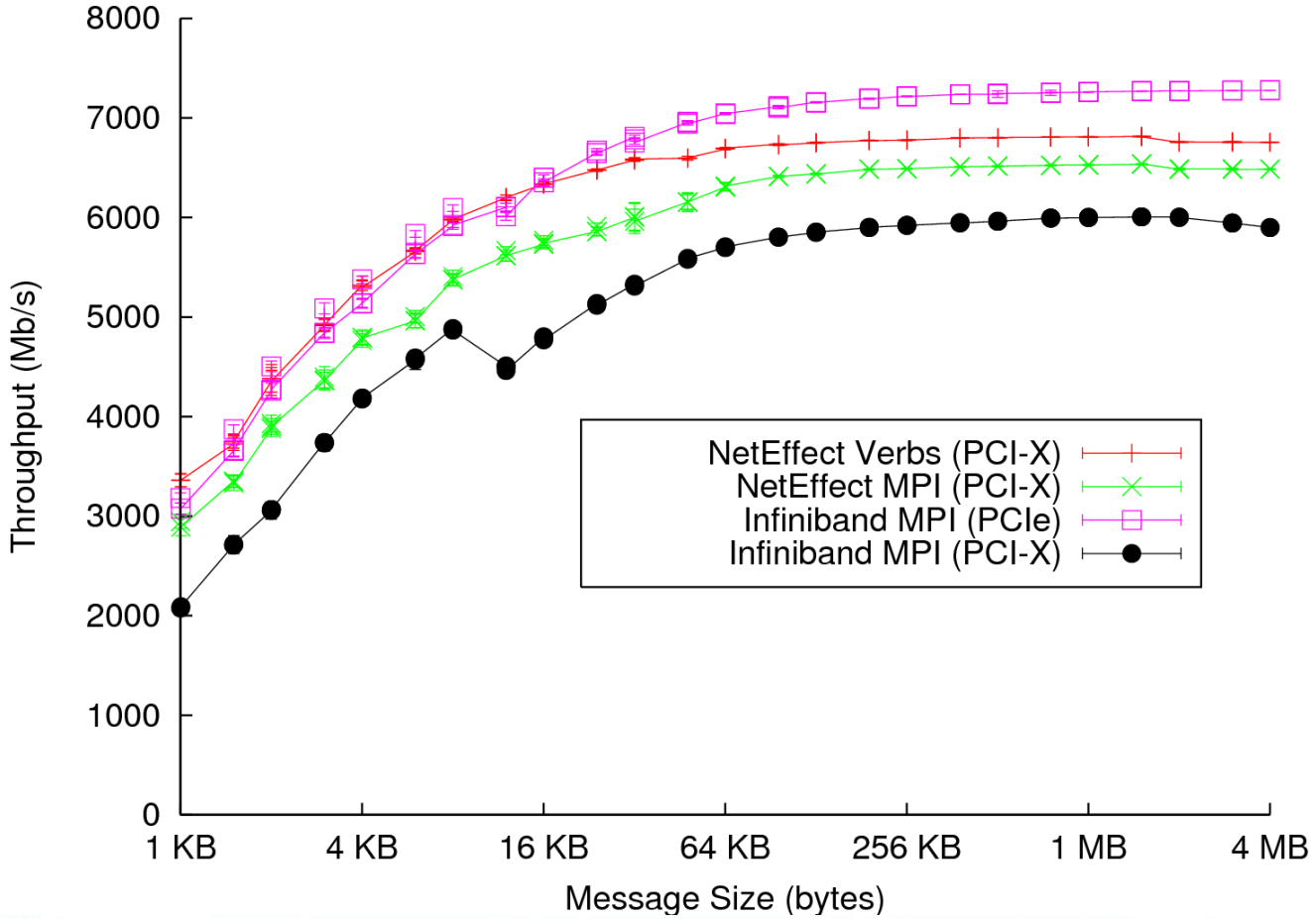
# PCI-X vs PCIe



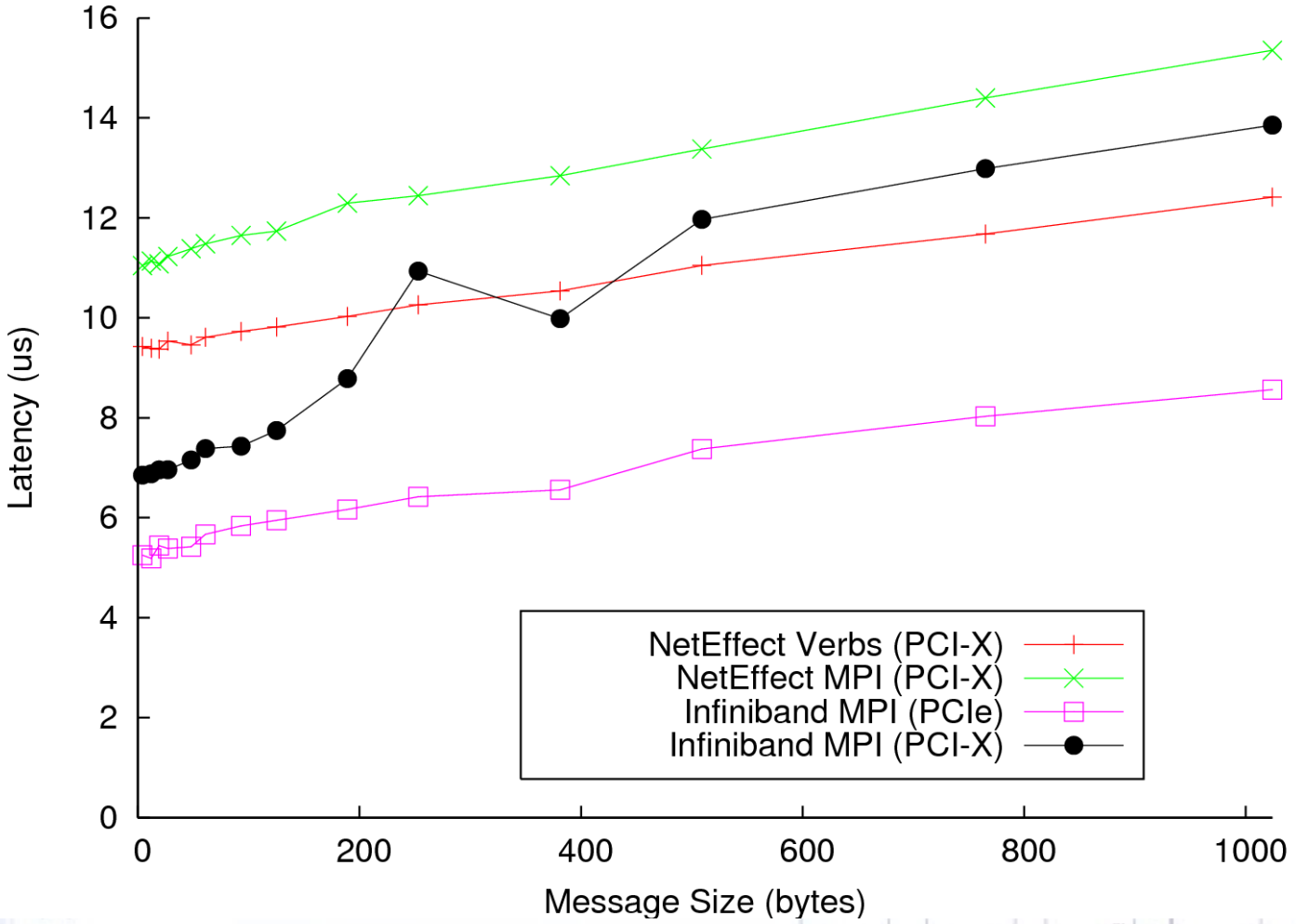

Interface limit: 16+16 Gb/s

Interface limit: 8.5 Gb/s

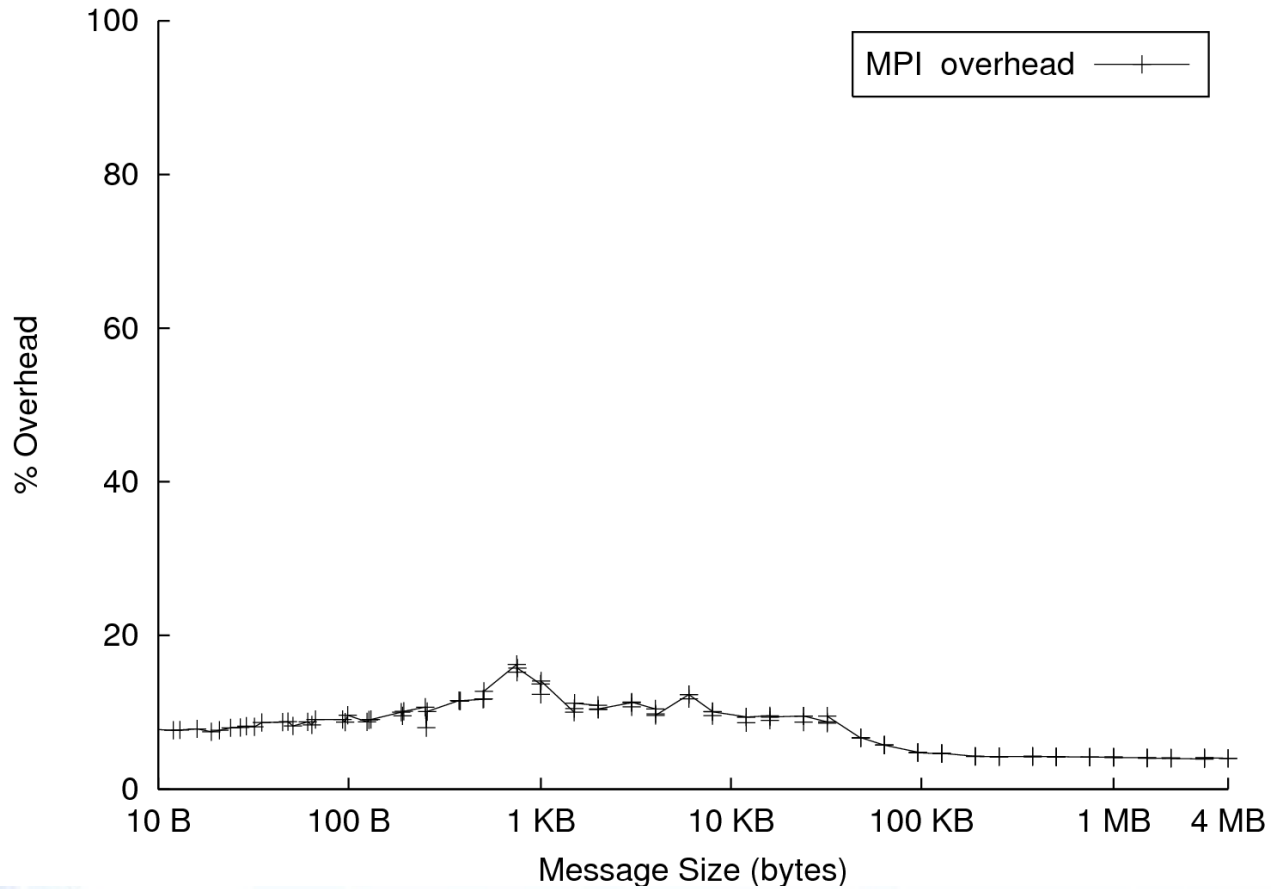
# Throughput


8000
7000
6000
5000
4000
3000
2000
1000
0
Throughput (Mb/s)
1 KB
4 KB
16 KB
64 KB
256 KB
1 MB
4 MB
Message Size (bytes)
NetEffect Verbs (PCI-X)
NetEffect MPI (PCI-X)
Infiniband MPI (PCIe)
Infiniband MPI (PCI-X)

# Latency

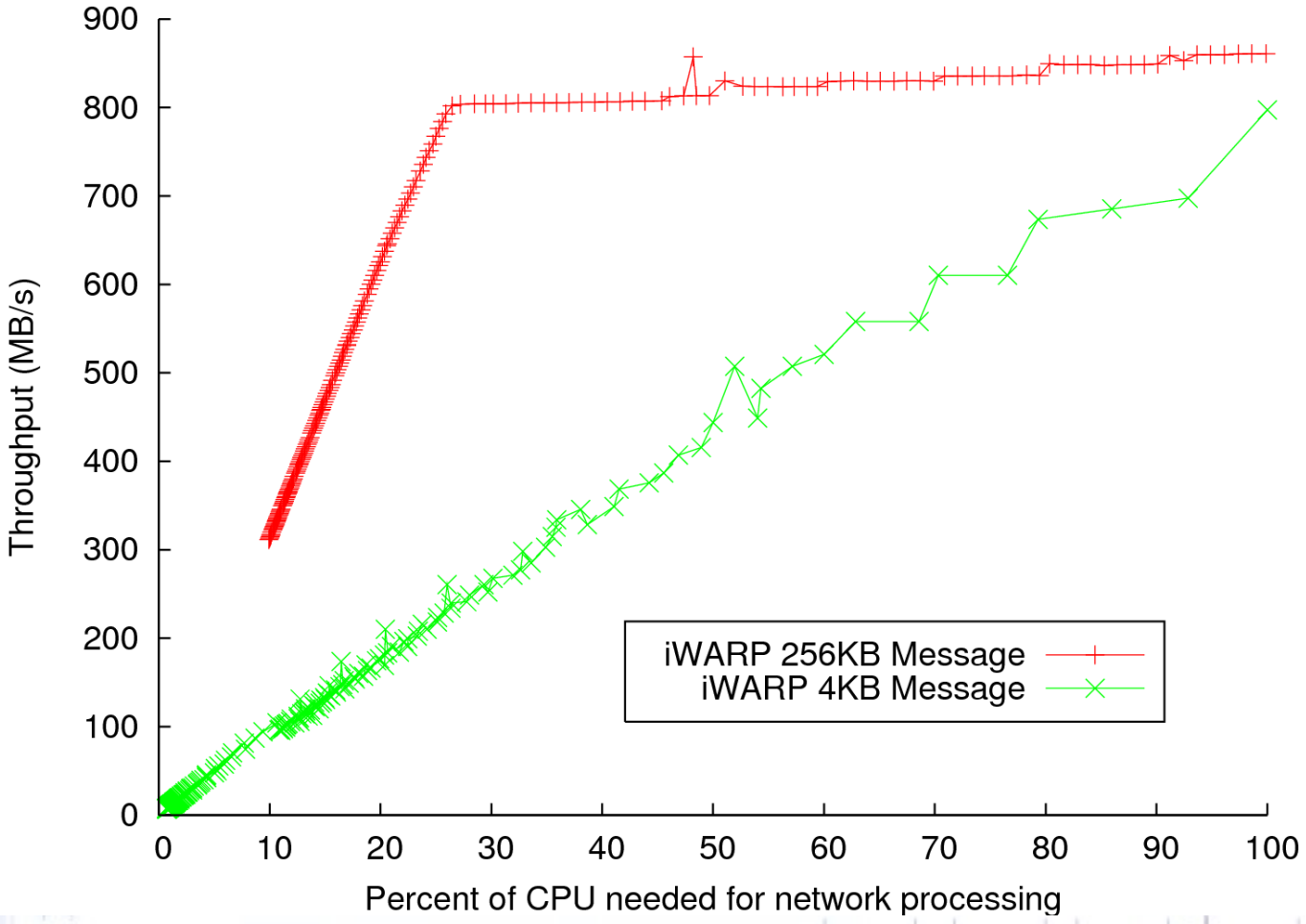

16
14
12
10
8
6
4
2
0
Latency (us)
0
200
400
600
800
1000
Message Size (bytes)
NetEffect Verbs (PCI-X)
NetEffect MPI (PCI-X)
Infiniband MPI (PCIe)
Infiniband MPI (PCI-X)

# MPI Overhead

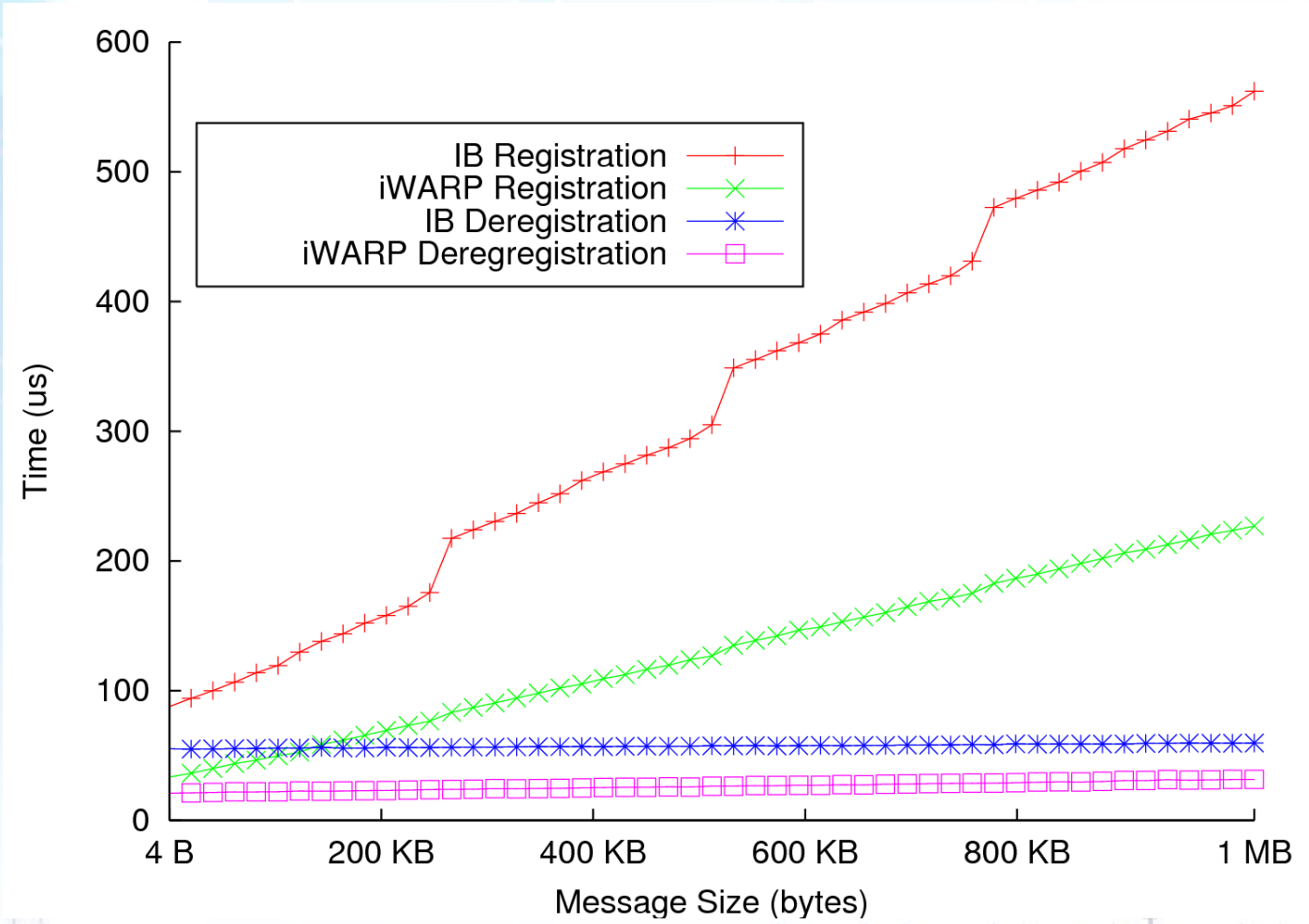



*In previous work: TCP overhead 20% and 1 Gig iWARP overhead 10%

# CPU Utilization



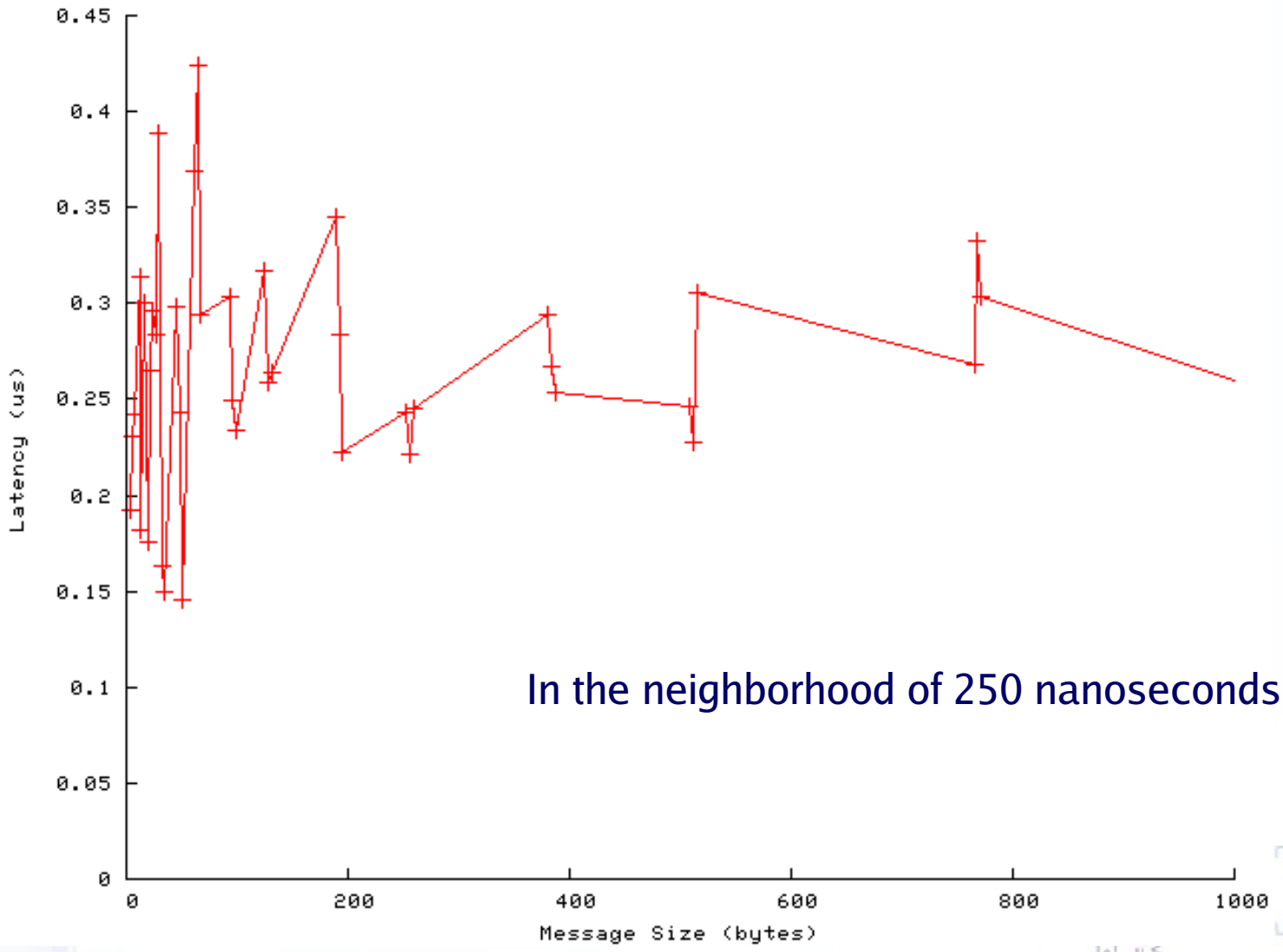
# Memory Registration



# Switch Overhead

- Ethernet switches are notorious for high latency
  - store and forward
  - on the order of a few microseconds
- Big selling point for InfiniBand
  - sub 250 nanosecond switch latency
- New cut through switch technology allows similar
  - Tested Fulcrum Micro and found similar latency

# Fulcrum switch latency




In the neighborhood of 250 nanoseconds

OSC

# Programming

- Verbs API
  - Just like IB (VAPI) each vendor has their own API
  - RDMAC verbs spec, Ammasso, NetEffect both use
    - Chelsio?
- MPI
- DAPL
- OpenFabrics (formerly OpenIB)
  - Common API for any RDMA device
  - iWARP uses RDMA CM
  - IB can as well (supposedly)

# Target Applications

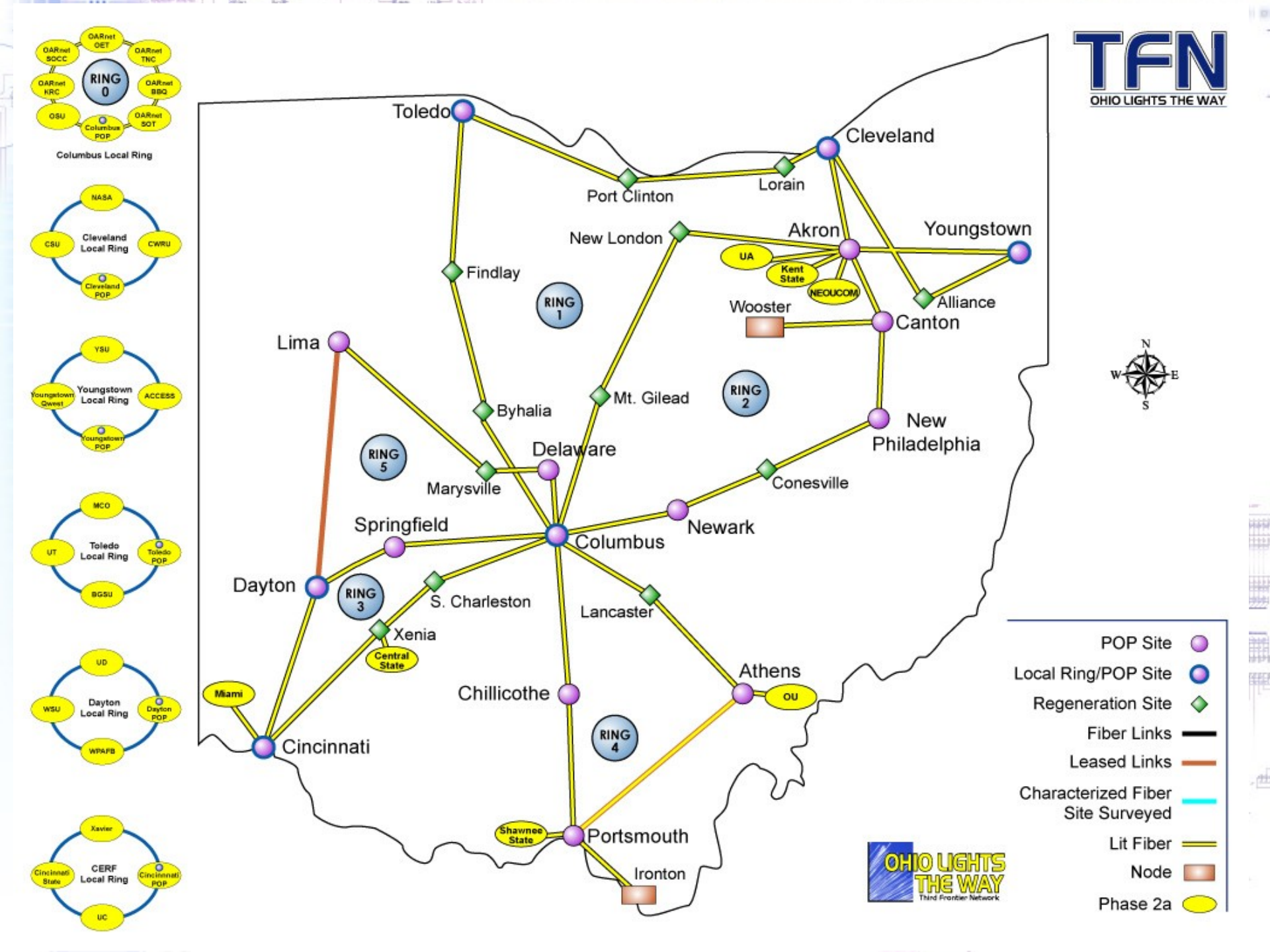
- Clustering
  - Even though IB has lower lat and higher bw
  - Majority of clusters (Top 500) are still Ethernet
    - Gigabit at that!
    - People unwilling to move away from familiar network despite 20X BW improvement
    - Price issues too

# Target Applications cont....

- Storage
  - NAS - Network Attached Storage
    - Benefits of RDMA from userspace
  - SAN - Storage Area Network
    - High throughput - low latency

# Target Applications cont...

- Wide Area Network
  - InfiniBand Not Applicable
    - Can not route IB
    - Can not span subnets
  - Internet is THE WAN
    - TCP/IP based
    - Most widely used “application”


TFN
OHIO LIGHTS THE WAY
RING 0
OARnet OET
OARnet TNC
OARnet BBQ
OARnet SOT
Columbus POP
OSU
OARnet KRC
OARnet SOCC
Columbus Local Ring
NASA
CSU
CWRU
Cleveland POP
Cleveland Local Ring
YSU
Youngstown Qwest
ACCESS
Youngstown POP
Youngstown Local Ring
MCO
UT
Toledo POP
BGSU
Toledo Local Ring
UD
WSU
Dayton POP
WPAFB
Dayton Local Ring
Xavier
Cincinnati State
Cincinnati POP
UC
CERF Local Ring
Toledo
Cleveland
Port Clinton
Lorain
New London
Akron
Youngstown
UA
Kent State
NEOUCOM
Findlay
RING 1
Wooster
Alliance
Canton
Lima
Byhalia
Mt. Gilead
RING 2
New Philadelphia
RING 5
Delaware
Marysville
Conesville
Springfield
Columbus
Newark
Dayton
RING 3
S. Charleston
Lancaster
Xenia
Central State
Miami
Athens
OU
Chillicothe
RING 4
Cincinnati
Shawnee State
Portsmouth
Ironton
N
W
E
S
POP Site
Local Ring/POP Site
Regeneration Site
Fiber Links
Leased Links
Characterized Fiber Site Surveyed
Lit Fiber
Node
Phase 2a
OHIO LIGHTS THE WAY
Third Frontier Network

# Current iWARP Work

- Software iWARP
  - Possible to emulate iWARP protocol in software
  - Direct benefits for the server
  - Indirect client benefits
    - Server is more responsive and able to handle more requests
  - OpenFabrics port (API)

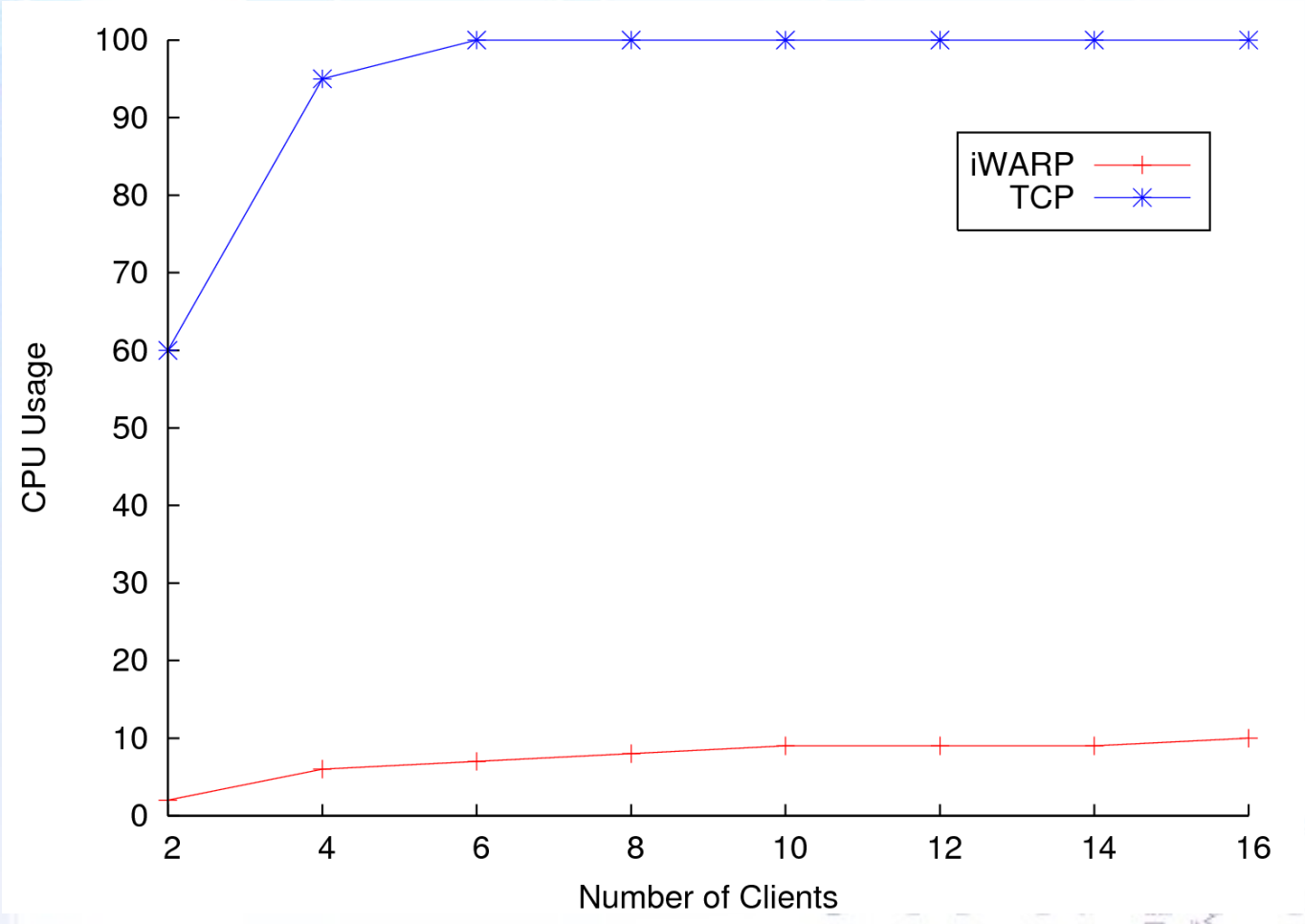
# Current iWARP work cont...

- Interoperability
  - Ammasso & SW iWARP - WORKS!
  - NetEffect & SW iWARP - Partially works.
    - Still got some kinks to work out
    - Issue is connection not data transfer
  - Ammasso & NetEffect
    - Working on getting hardware in place

# Current iWARP work cont...

- Apache RDMA Module (mod_rdma)
  - Works for Ammasso (ccil)
  - Porting to OpenFabrics
  - Subject of a poster accepted to SC|06
  - Subject of a demo in OSC's booth at SC|06

# mod_rdma Performance



# Future iWARP Work

- iWARP Storage systems
  - PVFS
- Parallel FTP
  - Already have simple client/server
  - Demo at SC|05
- Real world applications
  - Is 10  Gig enough right now?
  - Is 20Gig DDR IB really necessary?
  - Can TOE provide good enough performance?

# Thanks!

- Questions?
- Contact:
  - Dennis Dalessandro
  - dennis@osc.edu
  - http://www.osc.edu/~dennis/iwarp